

For reprint orders, please contact reprints@future-science.com

On the origins of three-dimensionality in drug-like molecules

Aim: Many medicinal chemistry-relevant structures and core scaffolds tend toward geometric planarity, which hampers the optimization of physicochemical properties desirable in drug-like molecules. As challenging drug target classes emerge, the exploitation of molecular three-dimensionality in lead optimization is becoming increasingly important. While recent interest has emphasized the importance of enhanced three-dimensionality in molecular fragment designs, the extent to which this is required in core scaffolds remains unclear. **Materials & methods:** Three computational methods, Scaffold Tree deconstruction, Synthetic Disconnection Rules retrosynthetic deconstruction and virtual library enumeration, are applied, together with the descriptors plane of best fit and principal moments of inertia, to investigate the origins of three-dimensionality in drug-like molecules. **Conclusion:** This study informs on the stage at which molecular three-dimensionality should be considered in drug design.

First draft submitted: 6 May 2016; Accepted for publication: 1 July 2016;
Published online: 30 August 2016

Keywords: drug-like molecules • library enumeration • molecular shape • molecular three-dimensionality • plane of best fit • principal moments of inertia • scaffolds • ternary density plots

The modulation and optimization of molecular properties are the essence of molecular design in medicinal chemistry. Many different parameters are convoluted in bringing a drug to market, neatly encapsulated in the safety and therapeutic efficacy of the drug product. The obvious properties such as potency against a biological target can be optimized in a variety of ways, but other characteristics of a given molecule may be subtly influenced by the modulation of an array of molecular properties through multiobjective optimization [1]. Recently, the three-dimensionality of chemical structures, from molecular fragments to drug-like molecules, has gained widespread attention due to reported evidence of key benefits to increased three-dimensionality in drug molecules [2–5].

An increase in the probability of clinical success has been implicated in the three-dimensionality of molecules through various descriptors including shadow indices [6] and the complexity descriptor, fraction of sp^3 -hybridized carbons [7]. More recently, it has also been shown that an increase in complexity using fraction of sp^3 -hybridized carbons can lead to a reduction in promiscuity and CYP450 inhibition [8].

The shape of chemical structures in drug discovery is a crucial component for evoking molecular recognition events with biological targets [9]. Natural products often contain enhanced three-dimensionality in their structures reflecting the interactions necessary with their biological partners [10]. Protein–protein interactions, a challenging target class that remains difficult to drug, are recognized as

Joshua Meyers¹,
Michael Carter¹, N Yi Mok^{*,1}
& Nathan Brown^{**,1}

¹Cancer Research UK Cancer Therapeutics Unit, Division of Cancer Therapeutics, The Institute of Cancer Research, London, SM2 5NG, UK

*Author for correspondence:
Yi.Mok@icr.ac.uk

**Author for correspondence:
Nathan.Brown@icr.ac.uk

**FUTURE
SCIENCE** part of

fsg

requiring enhanced three-dimensionality in the recently documented development of small molecule therapeutics. For example, the published inhibitors for BRD4/chromatin [11], LEDGF/p75 integrase [12], Bcl2/Bcl-xL family proteins [13] and the MDM2/p53 interaction [14], all exhibit enhanced 3D shape to complement their respective protein–protein interaction binding sites.

Physicochemical properties may also be enhanced to overcome particular challenges in drug design. Aqueous solubility of molecules is important for appropriate systemic exposure. Solubility may be improved by consideration of the 3D shape of molecules, with improvements in solubility achieved through disrupting the planarity of the molecular structures leading to disruption within solid state crystal lattice packing [15,16].

A number of molecular descriptors have been reported to characterize the three-dimensionality of molecular structures [17]. Here, we apply two recently published and widely adopted molecular descriptors to characterize three-dimensionality of molecular structures: principal moments of inertia (PMI) [18] and plane of best fit (PBF) [19]. While other molecular descriptors of three-dimensionality have been reported, we focus on these two as recently reported descriptors of relevance for our purposes.

The PMI descriptors provide a means of assessing the extent to which a given molecular geometry is rod

shaped, disc shaped and sphere shaped. The PMI are normally visualized on a ternary plot (Figure 1A) where the top-left vertex represents purely rod shaped, the vertex at the bottom represents those molecules that are completely disc shaped and the top-right vertex those structures that are entirely sphere shaped. Therefore, a molecular structure of particular shape will lie somewhere on the continuum between those three vertices representing the degree to which its morphology exhibits those primitive shape classes. A key characteristic of the normalized PMI descriptors is that there is no size dependence, therefore, for instance, adamantane and buckminsterfullerene exhibit identical normalized PMI ratios.

The PBF descriptor is a recent three-dimensionality descriptor that was designed, implemented and first reported in the literature by Firth *et al.* [19]. Given a particular molecular geometry, PBF identifies the plane of best fit running through that molecule that minimizes the distance of heavy atoms from the plane in Å. In this way, PBF is analogous to simple linear line-fitting in two dimensions, but extends its calculation to three dimensions. Once the plane of best fit has been determined, the PBF descriptor is calculated as the sum of the distances of the heavy atoms from the plane divided by the number of heavy atoms (Figure 1B). In contrast to the PMI descriptors above,

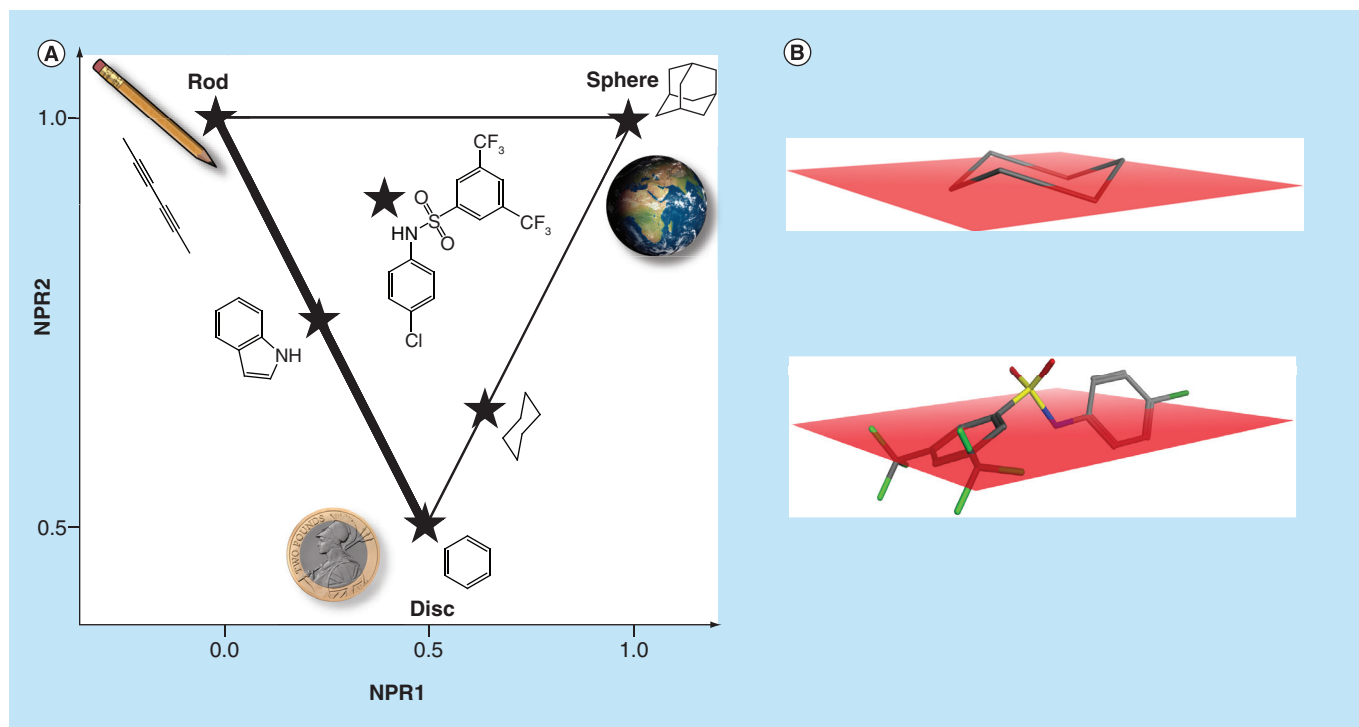


Figure 1. Schematic illustrations of the molecular descriptors used to characterize molecular three-dimensionality. (A) Schematic of the principal moments of inertia ternary plot with exemplar points elaborated with their respective chemical structures. (B) The plane of best fit for cyclohexane (above) and an exemplar drug-like molecule (below). The plane of best fit descriptor is the average distance of each heavy atom from the plane of best fit through all heavy atoms.

the PBF descriptor does exhibit size dependency in its calculation. Therefore, adamantane and buckminsterfullerene will have substantially different PBF values – 0.79 and 1.76 Å, respectively.

In this paper we report on the origins of three-dimensionality in drug-like molecules through analysis of more than one million chemical structures that have been published in the medicinal chemistry literature available from the ChEMBL 21 database [20]. The ChEMBL contains the chemical structures of compounds that have been tested and reported in experimental assays and are therefore considered inherently drug-like. Three methods were applied to investigate molecular substructures of these molecules, a well-established scaffold-based deconstruction technique, a previously reported retrosynthetic disconnection approach and a prospective analysis of new molecular structures enumerated *in silico*.

The first approach uses the published Scaffold Tree to systematically deconstruct molecules using the ring-focused disconnection rules implemented in the Scaffold Tree [21]. For a given molecule, pendant ring systems are pruned iteratively, generating different levels of the Scaffold Tree. At each level, the three-dimensionality of the scaffold can be evaluated and allows for the retrospective understanding of the possible origins of three-dimensionality in drug-like molecules.

In addition to the Scaffold Tree method, it is possible to use the recently reported Synthetic Disconnection Rules (SynDiR) protocol to deconstruct chemical structures. Applying SynDiR to the ChEMBL drug-like dataset renders chemically plausible substructures according to a prioritized list of retrosynthetic disconnection rules [22]. The SynDiR approach attempts to simulate a retrosynthetic analysis of an expert medicinal chemist. The three-dimensionality of the resultant molecular substructures can then be analyzed using the three-dimensionality descriptors described above.

Last, the virtual enumeration of new molecular structures can be used to understand how the three-dimensionality of the enumerated molecules may originate from the original molecular scaffolds. Simulating the prospective exploration of chemical space in medicinal chemistry when optimizing multiple parameters, the enumeration from scaffolds of varying inherent three-dimensionality allows us to investigate the contribution to three-dimensionality from the original scaffold versus those from the combination of the molecular substructures as an emergent phenomenon.

Using the two molecular descriptors of three-dimensionality and the three different approaches to investigate the exemplified drug-like chemistry space of interest, it is possible to systematically examine the origins of three-dimensionality in drug-like molecules. While

we are aware that chemical structures are often capable of adopting multiple conformations [23,24], hence likely affecting the associated descriptors of three-dimensionality, the study presented here applies a literature standard method that evaluates the 3D geometries of both molecules and their substructures using a single CORINA-derived conformation [25]. Through these structural analyses, guidance can be provided on the origins of three-dimensionality in drug-like molecules. Furthermore, a thorough understanding of the best approaches to introduce three-dimensionality will be of significant benefit to modern drug discovery and future medicinal chemistry design.

Materials & methods

Preparation of the ChEMBL database

From ChEMBL21 [26], 1,051,579 drug-like small molecules satisfying Lipinski's rule-of-five [27] with a minimum of one ring were used in this study. In order to compile this dataset, 1,560,490 unique structures from ChEMBL21 were annotated as 'small molecule'. Structures containing any macrocyclic ring (defined as ring size of 12 atoms or larger [28]) were then removed. After stripping salts and removing duplicates, 1,459,372 structures remained for further analysis. The remaining 1,438,214 structures were subjected to Lipinski Filters [27] as implemented using Pipeline Pilot [29]. The final dataset of drug-like small molecules contained 1,051,579 structures.

Conformer generation

CORINA [30] was used to generate coordinates for a single low-energy 3D conformation of each molecule in the ChEMBL21 dataset [31]. Hydrogens were added for structure generation and subsequently removed for all further analysis [25]. CORINA was employed with default parameters apart from the 'canon' flag, which was set to False. CORINA failed to generate 3D coordinates for 2411 molecules, over 90% of which were found to contain either chiral bridgehead atoms or a chiral center adjacent to a double-bonded π -system. Where molecules from ChEMBL21 contained undefined stereocenters, these were not enumerated to avoid biasing the dataset toward structures with multiple undefined stereocenters. Instead, CORINA selects an arbitrary stereoisomer unless the stereocenter forms part of a ring in which case the lowest energy conformation is selected.

Calculation of 3D descriptors

All 3D descriptors were calculated from a single CORINA-generated low energy conformer of each chemical structure with hydrogens removed. The PMI values were calculated using the built-in protocol

PMI implemented in Pipeline Pilot [29]. The normalized PMIs were calculated from the original *PMI_X*, *PMI_Y* and *PMI_Z* descriptors automatically in this component, yielding *NPR1* and *NPR2*. The PBF descriptor was calculated as implemented in Python as part of the RDKit API (version 2015.09.02) [32].

Generation of the Scaffold Tree

The curated ChEMBL21 dataset was analyzed using the Scaffold Tree method as implemented in the Molecular Operating Environment [33]. This analysis resulted in 3,211,860 individual levels from 1,051,579 unique molecular structures. The individual levels generated from each molecule were then opened in Pipeline Pilot [29] and the molecular structures generated from the SMILES strings, resulted in the failure of 220 levels. The remaining 3,211,640 levels were merged on their SMILES strings giving a total of 384,531 unique levels of the Scaffold Tree with their frequency of occurrence retained. A single low-energy conformation for each of these Scaffold Tree levels was generated using CORINA [30] and hydrogen atoms explicitly removed after generation – 98 levels failed conformer generation.

Molecular deconstruction using SynDiR

SynDiR was implemented using Python and the RDKit API (version 2015.03.1) [32]. SMILES strings for 1,051,579 chemical structures in the curated ChEMBL21 dataset were deconstructed into 3,304,410 substructures. Duplicate substructures were merged using the Merge Molecules component in Pipeline Pilot [29]. Disconnection points for each substructure were retained separately. A single low-energy conformation was then generated for each substructure using CORINA [30] – 12 substructures failed conformer generation.

Enumeration of virtual libraries

- Core Scaffolds: exemplar core scaffolds were selected from the Level 1 Scaffolds with attachment point annotations using the Scaffold Tree component as implemented in Pipeline Pilot [29]. Level 1 Scaffolds were binned into defined PBF bins of width 0.1 Å. The most frequent unique Level 1 Scaffold containing only two attachment points was selected for PBF bins ranging from 0.0 to 0.8 Å;
- Terminal substituents: the twenty most frequent terminal substituents containing only one attachment point from the unique SynDiR molecular substructures were selected (Supplementary Figure 1);

- Libraries enumeration: the virtual libraries were enumerated using the selected core scaffolds and terminal substituents applying the built-in protocol Enumerate using RGroups in Pipeline Pilot [29]. The twenty selected terminal substituents were attached to each of the two attachment points on each individual core scaffold, resulting in a total of 400 molecules per Level 1 scaffold.

Principal moments of inertia plots

Normalized PMI ratios were used to calculate rod-likeness, disc-likeness and sphericity as defined by Wirth *et al.* [34]. The three PMI of a molecular geometry are calculated through diagonalization of the moment of inertia tensor. These initial descriptors are normalized by dividing the two lower values by the highest value, yielding two mass-independent normalized PMI ratios, *NPR1* and *NPR2*, which can then be visualized in two dimensions – these normalized values are referred to as the PMI descriptors of a molecule. These were then used to plot ternary scatter and density plots implemented in Python making use of the python-ternary library [35]. An IPython Jupyter Notebook for creating ternary density plots is provided in the Supplementary Information.

Results & discussion

A concept that has persisted in the field of medicinal chemistry is that the coverage of chemical space tends toward more planar molecules at the expense of increased three-dimensionality. This trend can be explicated through appropriate descriptors of three-dimensionality; where more planar molecules are represented by the rod-shaped and disc-shaped edge of PMI space, and a PBF value closer to zero. Figure 2A shows the PMI plot for the entirety of the ChEMBL drug-like chemistry space illustrating that these molecules do indeed tend toward more planar geometries – those lying close to the rod-shaped and disc-shaped edge – with a marked preponderance of rod-shaped molecules. Similarly, an investigation of the PBF distributions for the same ChEMBL drug-like space (Figure 2B) illustrates that the median of the PBF distribution is approximately 0.6 Å, over a wider range from 0 to 2 Å. This clear over-representation at the lower end of the range of the PBF values suggests that while molecular structures with enhanced three-dimensionality are possible and have been investigated, the majority of chemical structures reduced to practice through synthesis have a tendency away from more 3D space.

Analyses of both the PMI and PBF distributions of three-dimensionality from the ChEMBL drug-like dataset illustrate that, while enhanced 3D shape has

been reported in the literature, these chemical structures tend to be under-represented (Figure 2). While it is possible to speculate on reasons for this observed lack of 3D structures, such as conventions of synthesis or biological necessity [36], it is impossible to truly isolate the reasons for this observed imbalance. However, it is possible to investigate further the origins of three-dimensionality in drug-like molecular structures through observation of structures that exist and those that could exist through unbiased virtual library enumeration.

The parent molecule analysis of the chemical structures in the ChEMBL drug-like dataset, summarized in Figure 2, demonstrates a general lack of shape diversity and three-dimensionality. However, this does not inform on the origins of three-dimensionality in drug-like molecules, rather the ultimate designs that are synthesized. Therefore, the question remains open: from where does three-dimensionality originate in drug-like molecules. Using the Scaffold Tree algorithm [21] it is possible to apply systematic disconnection rules to pare back the parent molecules to their constituent scaffolds, ring-by-ring. The application of the Scaffold Tree algorithm to the parent molecules in the ChEMBL drug-like dataset is presented with both the PMI (Figure 3A) and PBF (Figure 3B) distributions.

The ternary density plots of the PMI values (Figure 3A) illustrate the density of individual scaffolds from level 8 (where present) down to level 0 in decreasing

number of rings, where level 0 scaffolds contain a single ring by definition. The color key conveys the density of individual scaffolds in each bin of the ternary plot, and is logarithmically scaled (note the dramatic frequency changes between each level when interpreting these figures). The higher levels of the Scaffold Tree are sparse since there are relatively few individual scaffolds with many rings (Figure 3C displays the number of scaffolds generated for each level of the Scaffold Tree for the entire ChEMBL drug-like dataset). As the Scaffold Tree levels are reduced, shifts toward more planar regions of PMI space can be observed. This reduction in overall three-dimensionality is even more pronounced in the box-and-whisker plot of the PBF distributions for the individual Scaffold Tree levels when compared with the original parent molecules (Figure 3B). The general trend is consistently toward more planar scaffolds as the size of the scaffolds is reduced.

In a recent publication on the scaffold diversity of exemplified medicinal chemistry [37], it was observed that Level 1 of the Scaffold Tree typically represents an appropriate objective and invariant scaffold definition. With this definition in mind, it can be observed clearly that there is a substantial reduction in three-dimensionality when a drug-like molecule is reduced to its molecular scaffold, in particular between the Level 2 and Level 1 scaffolds where the median value of PBF distributions decreases from 0.37 Å for Level 2 scaffold

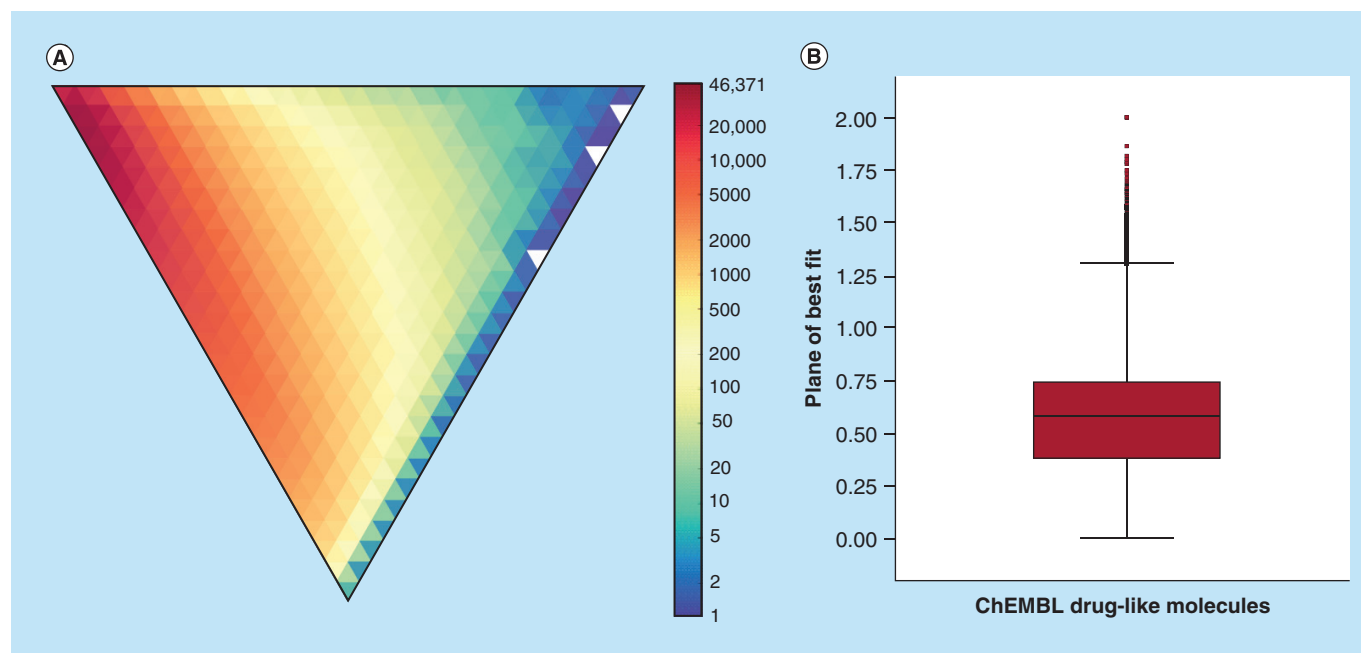


Figure 2. The three-dimensionality profile of the ChEMBL drug-like chemistry space analyzed throughout this study ($n = 1,045,172$). (A) The PMI ternary density plot and (B) the PBF box-and-whisker plot (in Å). The color key conveys the density of molecules in each bin of the ternary plot.

PBF: Plane of best fit; PMI: Principal moments of inertia.

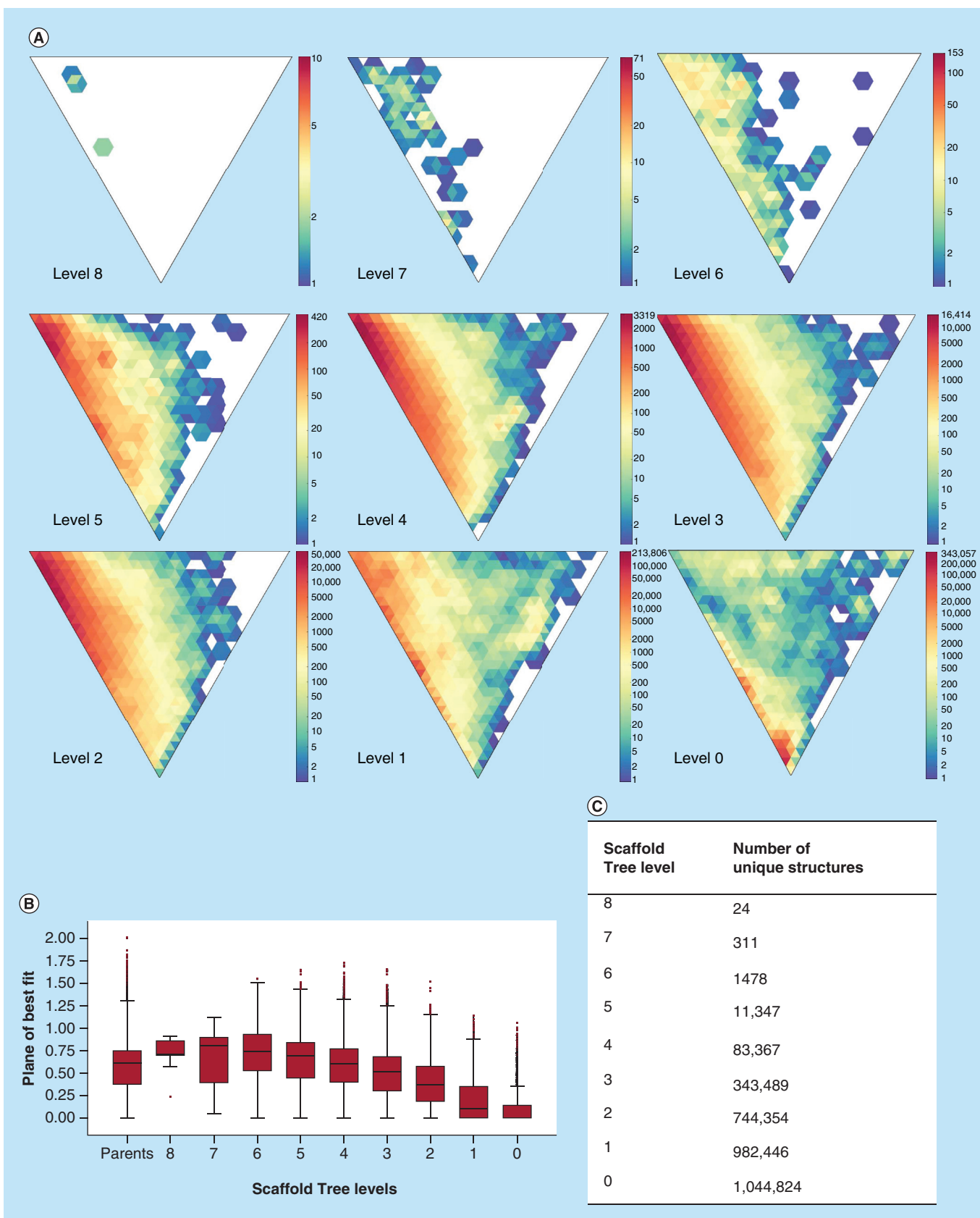


Figure 3. The three-dimensionality profile of the individual Scaffold Tree levels (see facing page). (A) The PMI ternary density plots and (B) the PBF box-and-whisker plot (in Å). The color key conveys the density of individual scaffolds in each bin of the ternary plot. The table in (C) shows the number of structures contained within each Scaffold Tree level. (B) also includes the PBF box-and-whisker for the parent structures from ChEMBL for comparison; the equivalent PMI plot is shown in Figure 2A. PBF: Plane of best fit; PMI: Principal moments of inertia.

folds to 0.11 Å for Level 1 scaffolds (Figure 3B). This observation implies that many drug-like molecules in ChEMBL have been derivatized from molecular scaffolds that are relatively planar. Another striking observation is the shift from rod-shaped to disc-shaped structures between the Level 1 and level 0 scaffolds in the Scaffold Tree. These phenomena can be explained by the prevalence of bicycles and monocycles in these two levels, respectively.

Using the Level 1 scaffold definition, it is possible to identify exemplified patterns of derivatization using their representatives from the ChEMBL drug-like dataset. This analysis can be used to understand the extent to which the three-dimensionality of the scaffolds may contribute to the overall three-dimensionality of the parent molecules. Figure 4 illustrates eight exemplar core scaffolds with increasing three-dimensionality selected as the most frequent disubstituted Level 1 scaffolds from eight uniformly divided PBF bins, ranging from 0 to 0.8 Å. The most frequent disubstituted Level 1 scaffolds in the higher PBF bins above 0.8 Å represent only a small number of parent molecules, hence those were not further analyzed. The ternary plots and box-and-whisker plots show the PMI and PBF distributions for each core scaffold, respectively. The ternary plots report the PMI of the core as a black star and every parent molecule containing that core as red points. The PBF values are reported in the box-and-whisker plots with the PBF value of the core on the left and the PBF distributions of the parents containing those cores on the right.

Figure 4 demonstrates that it is possible to modulate the three-dimensionality of parent molecules designed around a common scaffold regardless of the local three-dimensionality of the core scaffold itself, albeit to different degrees. Cores A and B, for example, while representing the two most planar scaffolds, are present in parent molecules that show broad sampling of PMI space. The rod-shaped tendency of the parent molecules from cores C and E can be partially explained by the linear substitution patterns of those scaffolds; this derivatization pattern has been recently shown to promote rod-likeness in PMI space [36]. Core D, although arguably a reasonable medicinal chemistry starting point, exhibits poor exploration of 3D space, tending mainly toward rod-shaped space, and not exploring toward the sphere-shaped vertex at all. Furthermore, all reported parent molecules containing Core F, an example of a sphere-shaped bridged ring scaffold,

represent a relative reduction in their sphericity. This observation highlights the potential limit of the exploration of 3D shape space of an exemplar-bridged scaffold that may otherwise be expected to promote three-dimensionality in exemplified derivatives. Across all PBF bins, the median of the PBF distributions for the parent structures can be seen to converge to a similar range of values irrespective of the PBF value of the originating core scaffold.

This retrospective experiment analyzing exemplified core scaffolds and their derivatized parents from the ChEMBL drug-like dataset suggests that, based on our analysis, there is no clear rationale for increasing the complexity or three-dimensionality of core scaffolds for an enhancement in the three-dimensionality of the parent molecules. It is the connectivity of the substructures that appears to contribute the most to the three-dimensionality of the parent molecules, rather than the local three-dimensionality of the core substructures themselves. Therefore, it can be concluded from this analysis that 3D molecules can be constructed from 2D core scaffolds and substructures.

While the retrospective experiment exemplified in Figure 4 provided some plausible insights into the origins of three-dimensionality of existing molecules from the ChEMBL drug-like dataset, it does represent a biased dataset subject to the requirements of different drug design projects at different times and the extent that a core scaffold has been explored synthetically. However, using the common core scaffolds selected in that experiment, an objective study can be devised that enumerates virtual libraries for each of the core scaffolds with the same substituents. To achieve this, it is necessary to identify common terminal substituents present in drug-like molecules. Therefore, using the ChEMBL drug-like dataset, the set of substructures present has been mined using a recently published set of retrosynthetic disconnection rules, SynDiR [22]. This set of SynDiR-generated substructures can also be assessed in parallel to the Scaffold Tree output, reducing possible biased observations from a single deconstruction algorithm.

Figure 5 illustrates the three-dimensionality of the SynDiR-generated substructures derived from the ChEMBL drug-like dataset. Figure 5A shows the ternary density plot representing the PMI values of the actual frequency of occurrence of each substructure in the dataset, while Figure 5B shows the unique occurrence of each substructure with all duplicates removed.

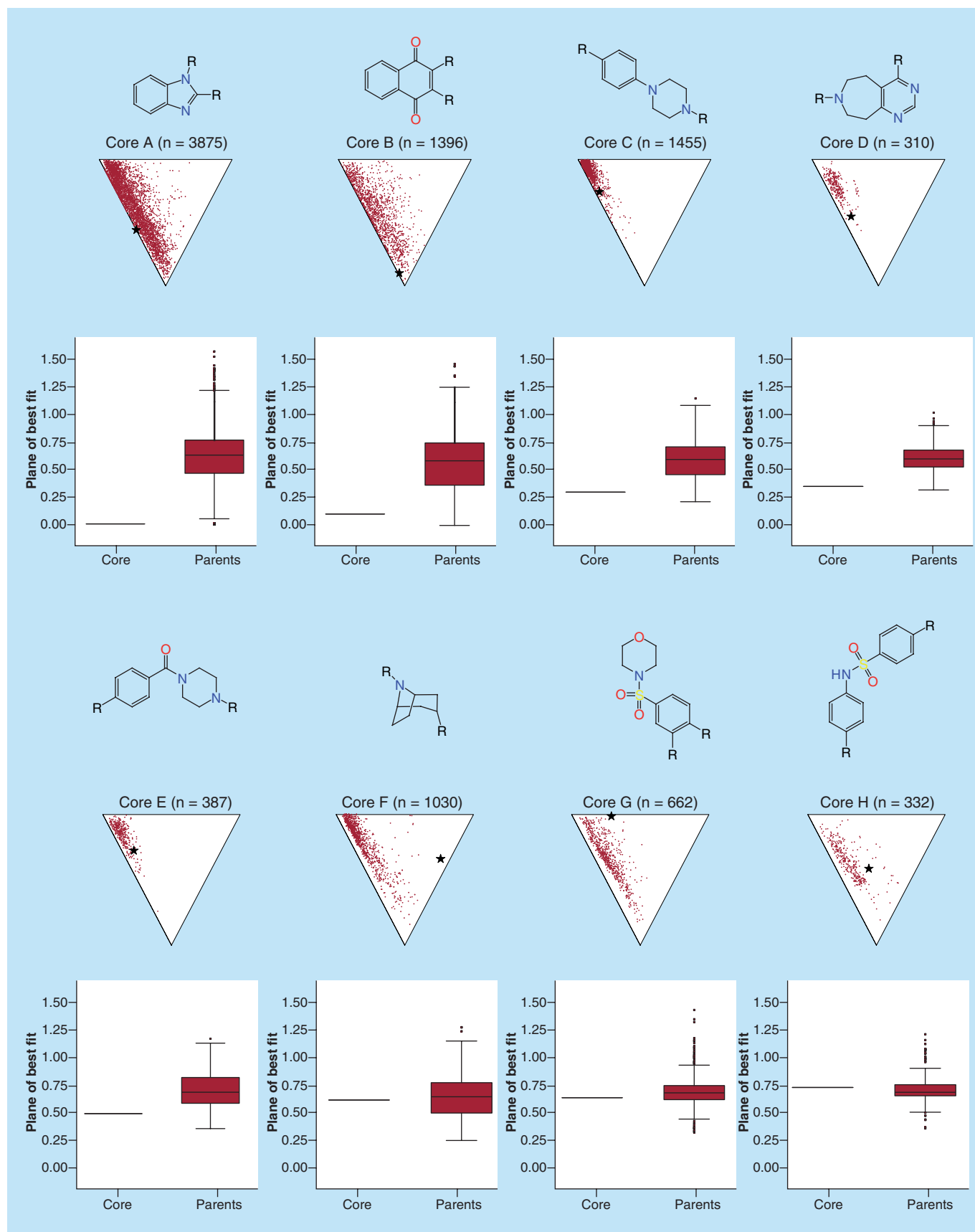


Figure 4. Exemplified Level 1 scaffolds (black star in principal moments of inertia ternary plots) of increasing 3D character and their existing parent molecules in ChEMBL (red points in principal moments of inertia ternary plots) plotted in principal moments of inertia ternary plots and plane of best fit box-and-whisker plots (in Å) (see facing page). R in structures denote substitution positions and are not necessarily identical substituents.

From both figures it is possible to recognize that there is a prevalence of substructures that are inclined toward the rod-shaped to disc-shaped edge and therefore exhibit strong planarity. When considering the frequency of occurrence (Figure 5A) and the unique occurrences (Figure 5B), the substructure morphology can be seen to shift toward enhanced three-dimensionality. This observation highlights that, while there is precedence for more 3D substructures, these are not represented well in their actual occurrence in exemplified medicinal chemistry. This observation agrees with previous evidence reported here from the Scaffold Tree analysis (Figure 3) that suggests, while substructures with enhanced three-dimensionality are present in the output of synthetic medicinal chemistry, they are not highly represented in the actual chemical structures synthesized and reported in the literature.

The PBF values for the SynDiR-generated substructures are given in Figure 5C as a 2D density heat map, with the logarithm of the actual frequency of occurrence on the x-axis and the PBF on the y-axis. This figure reinforces the conclusion from Figure 5A that there are many different planar substructures and that they occur very frequently in medicinal chemistry space whereas 3D substructures with high PBF values are sparsely represented. Similarly, Figure 5D illustrates the relationships between the actual frequency of occurrence of substructures, the unique substructures and the original parent structures from the ChEMBL drug-like dataset. This box-and-whisker diagram reiterates the prevalence of planar substructures that are commonly assembled in medicinal chemistry even though available 3D substructures have been sampled.

Analysis of the constituent substructures used in medicinal chemistry exemplified in the literature demonstrates that most of the chemistry space covered tends away from three-dimensionality in practice. Given the analysis of the scaffolds, in particular the Level 1 scaffolds, and the constituent substructures exemplified, it is now possible to repeat the experiments reported in Figure 4. However, rather than investigating the three-dimensionality of core scaffolds and their exemplars in the ChEMBL drug-like dataset, the most common core scaffolds of defined three-dimensionality, together with the most common terminal substituents, were used to objectively enumerate a virtual library containing core scaffolds of increasing three-dimensionality. Figure 6 presents enumerated virtual libraries centered on the same set of disubstituted Level 1 scaffolds analyzed in the previous

experiment. The enumerated libraries use the twenty most common terminal substituents from the substructures generated using the SynDiR retrosynthetic algorithm (Supplementary Figure 1); therefore the resultant virtual libraries vary only in the scaffold used in each case. The PBF values for the twenty terminal substituents were very low, ranging from 0 to 0.25 Å, reflecting the largely 2D nature of the most frequently occurring substructures identified in Figure 5. This study allows for the investigation of the contribution to three-dimensionality of the core scaffolds. Although enumerated structures may not be drug-like or even synthesizable, they represent an objective method with solid foundation for the investigation of the origins of three-dimensionality in chemistry space.

In many respects, the PMI plots and PBF distributions shown for each core scaffold in Figure 6 mirror those in Figure 4. This is reassuring and shows the capability of virtual library enumeration to explore potential chemical space surrounding these cores. The more planar core scaffolds in Figure 6 (Cores A and B) exhibit the ability to sample broad 3D shape space, supporting previous observations that 3D molecules can be constructed from 2D core scaffolds and substructures (Figure 4). While the majority of the cores contribute positively to the three-dimensionality of the enumerated libraries, this increase is much more pronounced in the planar cores. This may in part be explained by the relative change in molecular size when substituents are attached to the core. While the terminal substituents have been kept constant in this enumeration experiment, the relative change in molecular size between the core scaffold and the enumerated structures is larger for the planar core scaffolds of lower MW (for instance, MW of Core A is 118.14 Da) than the more 3D core scaffolds (Core H, MW 233.29 Da).

More surprisingly, the enumerated library from Core F displays a more localized coverage of PMI space in comparison to that in Figure 4, with the enumerated library lacking coverage toward the disc-shaped vertex of PMI space. This difference might be attributed toward the more diverse substituents in exemplified chemistry space than those utilized to create the enumerated library. Smaller substituents in combination with the linear substitution pattern of this core may be responsible for the preponderance of rod-shaped molecules in the enumerated library. Nonetheless, in both cases Core F shows limited exploration of enhanced 3D shape space, once again highlighting the potential limit of exploration for an exemplar bridged

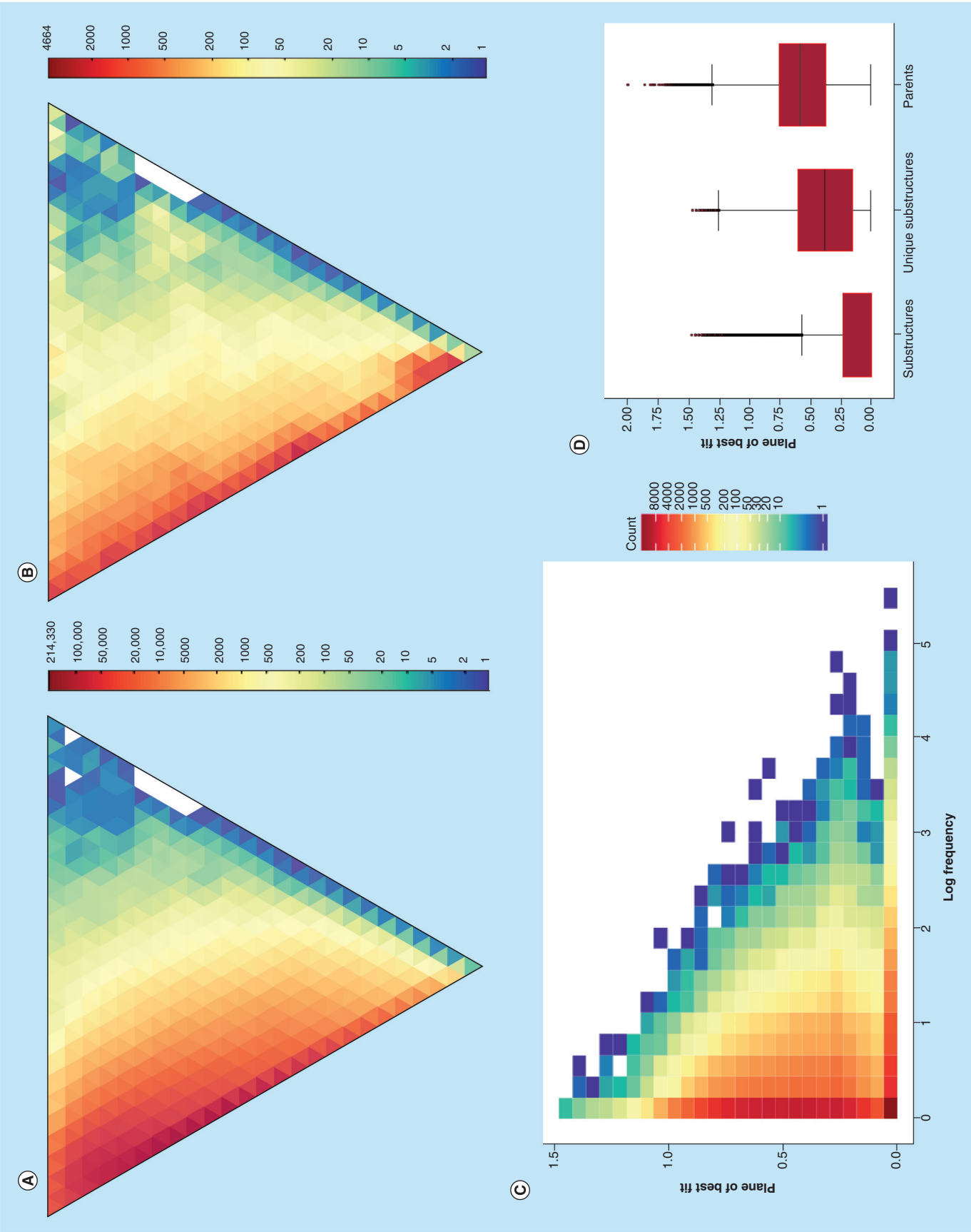


Figure 5. The three-dimensionality profile of Synthetic Disconnection Rules substructures (see facing page). (A) PMI ternary density plot showing the distribution of all SynDiR substructures and their frequency of occurrence (i.e., each substructure is counted the number of times it appears in the entire dataset) in ChEMBL ($n = 2,929,923$), and (B) the distribution of SynDiR substructures and their unique frequency of occurrence (i.e., each specific substructure is counted only once) in ChEMBL ($n = 170,758$). Figure (C) shows a heat map of, on the x-axis, the logarithm of the frequency of occurrence of SynDiR substructures in ChEMBL and, on the y-axis, their PBF values (in Å). The color of the heat map in figure (C) indicates the number of different substructures that occupy that particular bin, that is, they are similarly frequent and similar in 3D character, but are different structures. The color key conveys the density of individual substructures in each bin of the plots. Figure (D) is a box-and-whisker plot of all SynDiR substructures, the unique SynDiR substructures and the parent molecules in ChEMBL, against their respective PBF values (in Å). PBF: Plane of best fit; PMI: Principal moments of inertia; SynDiR: Synthetic Disconnection Rules.

scaffold that may otherwise be expected to promote three-dimensionality.

While the intrinsic three-dimensionality of the core scaffolds does not appear to contribute extensively to the PMI shape space coverage of the enumerated libraries, comparisons of the trend in median values of the PBF distributions suggest more 3D core scaffolds could result in more 3D enumerated libraries. Although this observation implies that more 3D core scaffolds may represent a good starting point in accessing molecules of enhanced three-dimensionality, it is noteworthy that the range of the PBF distributions decreases as the three-dimensionality of the core increases. This trend recapitulates the previous analysis of exemplified structures in ChEMBL (Figure 4), which shows that exploration of three dimensionality, in practice, can equally be accessible from planar core scaffolds.

The work presented here offers a range of interesting insights into the origins of three-dimensionality in medicinal chemistry-relevant structures. Early in a drug design project, it is important to have a clear motivation and design ethic prior to the introduction of more complex 3D cores. One of the main reasons to consider carefully the introduction of complex 3D cores is the inherent challenge in their synthesis and derivatization. Careful thought is required with regard to the potential improvements, such as physicochemical properties and improved intellectual property position, at the cost of effective exploration and exploitation of the chemistry space accessible through available synthetic methodologies. The probability of finding leads may also decrease as the molecular complexity of the core increases [38]. Similarly, the choice of substituents is important and the results presented here suggest that the combination of relatively planar scaffolds and substituents can lead to an effective sampling of 3D shape space.

Conclusion & future perspective

The benefits afforded by increased three-dimensionality in drug-like molecules, such as potency and enhanced physicochemical properties, have attracted much attention toward the exploration of novel 3D scaffolds and fragments. However, the origins of three-dimensionality in drug-like molecules have remained

largely unexplored. Here we present an investigation of the origins of the 3D nature of drug-like molecules by studying structures that exist in the ChEMBL database, and those that could exist through unbiased virtual library enumeration.

Inspection of the three-dimensionality, measured by both PMI and PBF, exhibited by molecules in the ChEMBL drug-like dataset demonstrates that while a wide range of three-dimensionality has been reported in the literature, there is an under-representation of chemical structures displaying enhanced three-dimensionality.

Deconstruction of parent molecules according to the Scaffold Tree algorithm highlights a trend toward planarity as pendant ring systems are iteratively removed. This trend is recapitulated when parent molecules are retrosynthetically deconstructed as defined by SynDiR, which exposes the propensity of medicinal chemistry to repeatedly incorporate largely planar substructures. Both deconstruction methods consistently demonstrated that, while there is precedence for substructures of enhanced three-dimensionality, these are not proportionally represented by their actual occurrence in exemplified medicinal chemistry.

Using Level 1 scaffold definitions to investigate exemplar core scaffolds and their parent molecules in ChEMBL, analysis of the PMI and PBF distributions associated with each core scaffold revealed that it is possible for exemplified chemical structures to exhibit enhanced three-dimensionality regardless of the intrinsic three-dimensionality of the core scaffold. This has implications as this analysis suggests that 3D molecules can be constructed from more planar core scaffolds. It was also highlighted that the substitution pattern of the core scaffolds can have a decisive influence on the coverage of PMI space, as illustrated by exemplar scaffolds of linear and adjacent disubstitutions. Attempts to further minimize any potential bias from our analysis of exemplified synthetic medicinal chemistry space were also investigated. Virtual library enumeration was employed as an unbiased method of probing chemical shape space accessible from core scaffolds of increasing three-dimensionality. These enumerated libraries illustrate that while core scaffolds of greater three-dimensionality may offer good starting

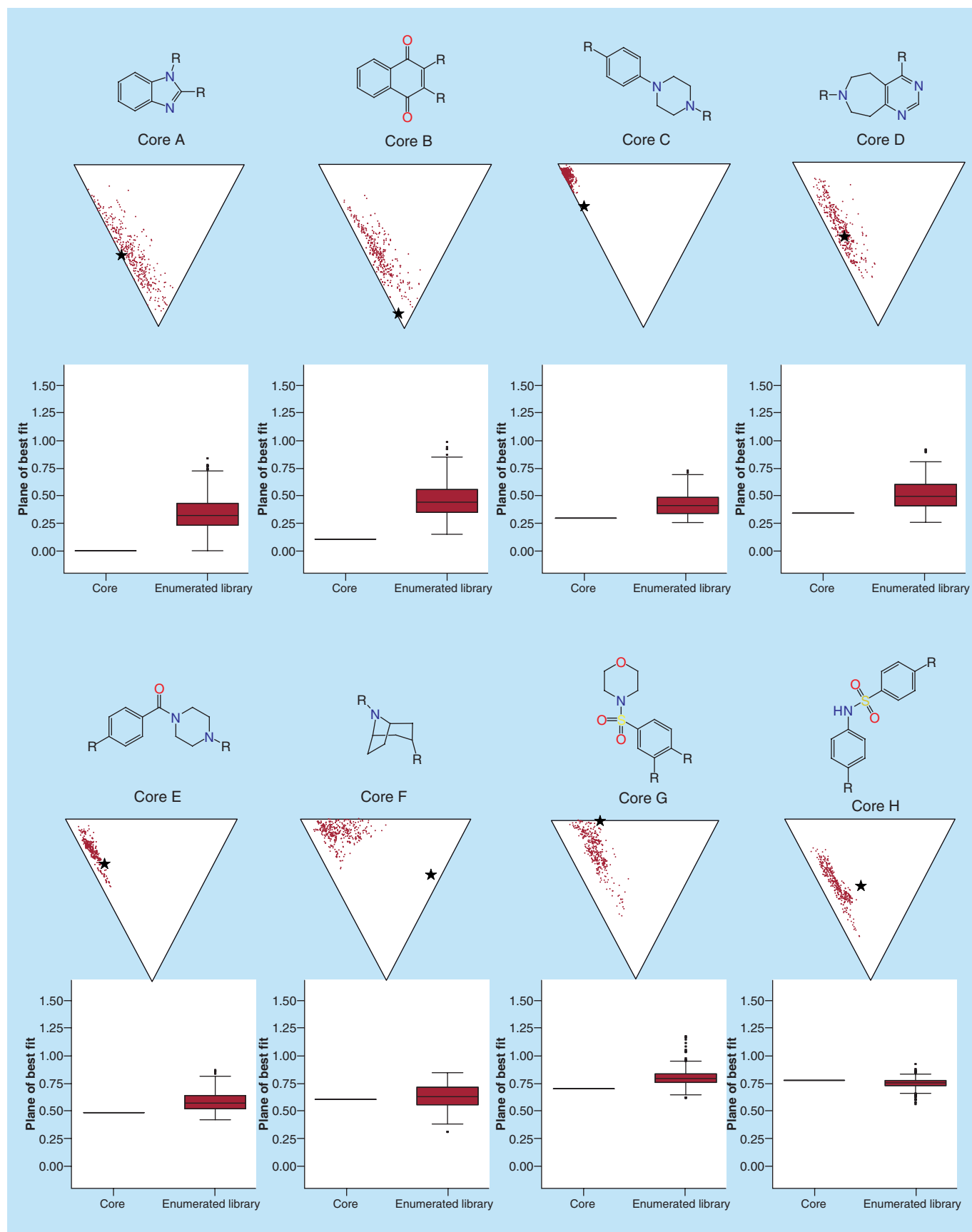


Figure 6. Exemplified Level 1 scaffolds (black star in principal moments of inertia ternary plots) of increasing 3D character and the enumerated libraries (n = 400, red points in principal moments of inertia ternary plots) using commonly observed medicinal chemistry substituents, plotted in principal moments of inertia ternary plots and plane of best fit box-and-whisker plots (in Å) (see facing page). R in structures denote substitution positions and are not necessarily identical substituents.

points for accessing 3D molecules, planar core scaffolds are comparably capable of sampling enhanced 3D shape space.

In conclusion, our study of ChEMBL drug-like molecules and their systematically deconstructed substructures provides some interesting insights into the origins of three-dimensionality in medicinal chemistry space. The results presented here suggest that relatively planar scaffolds and substituents can be combined to access a reasonable sampling of 3D shape space. While core scaffolds of enhanced three-dimensionality could facilitate sampling of more 3D shape space, we recommend that the consequences regarding synthetic accessibility and chemical space sampling of increasing 3D complexity be carefully considered before introducing more synthetically challenging 3D cores as a method of increasing molecular three-dimensionality.

Future perspective

While the potential benefits afforded by increased three-dimensionality in drug-like molecules are apparently multitudinous, there has been limited availability of objective methods evaluating molecular three-dimensionality that are easily interpretable when designing molecules. Using the simple descriptors PMI and PBF, this analysis has highlighted the prevalence of planar scaffolds and substructures that are exemplified in drug-like molecules reported in the medicinal chemistry literature regardless of the molecular three-dimensionality of the designed mol-

ecules. The observations in this analysis suggest that relatively planar scaffolds and substituents can be combined to access a reasonable sampling of 3D shape space. This is encouraging since many of the synthetic methodologies available to medicinal chemists are based on aromatic heterocyclic chemistry. However, the lack of well-represented scaffolds and substructures with intrinsic enhanced three-dimensionality in drug-like molecules analyzed may reflect the constraints in synthetic accessibility of these inherently more complex building blocks. The demand for novel synthesis in practice would become increasingly pressing as medicinal chemistry and drug discovery efforts diverge from a focus on protein kinase inhibition to emerging biological target classes including epigenetic targets and protein-protein interactions. The development of systematic data mining of historical synthetic records in medicinal chemistry laboratories will assist in unveiling lesser explored synthetic methodologies that may be enabled. As innovative methodologies and technologies continue to emerge in the medicinal chemistry synthetic toolkit, the diversity of synthetic tools available to medicinal chemists may enable and encourage the construction of designed drug-like molecules from more 3D scaffolds and substructures.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/full/10.4155/fmc-2016-0095

Executive summary

Existing drug-like molecules

- Existing drug-like molecules exemplified in the ChEMBL database tend toward planarity.
- However, a number of molecules are represented in more 3D space.
- 3D molecules are reported but are under-represented compared with more planar structures.

Scaffold Tree analyses

- Molecules pared back ring by ring using the Scaffold Tree show substantial reduction in three-dimensionality.
- Medicinal chemistry-relevant Level 1 scaffolds are typically planar.
- More planar scaffolds tend to permit greater modulation of molecular three-dimensionality.

Retrosynthetic analyses

- Retrosynthetic analysis conducted using Synthetic Disconnection Rules to deconstruct molecules to synthetically relevant substructures.
- Most substructures tend toward planarity.
- More 3D substructures do exist, but are used very rarely in practice.

Virtual library enumeration

- Virtual libraries can be enumerated from previously identified medicinal chemistry-relevant scaffolds and derived substructures.
- Greater modulation in three-dimensionality is possible when starting from planar scaffolds.
- However, more 3D scaffolds can facilitate sampling of more 3D shape space.

Financial & competing interests disclosure

This work and the authors have been supported by funding from the Wellcome Trust and Cancer Research UK. J Meyers is supported by Wellcome Trust grant 102361/Z/13/Z. M Carter, NY Mok and N Brown are supported by Cancer Research UK, grant C309/A11566. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject

matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

Open access

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 Unported License.

References

Papers of special note have been highlighted as:

• of interest; •• of considerable interest

- Nicolaou CA, Brown N. Multi-objective optimization methods in drug design. *Drug Discov. Today Technol.* 10(3), e427–e435 (2013).
- Kumar A, Voet A, Zhang KYJ. Fragment based drug design: from experimental to computational approaches. *Curr. Med. Chem.* 19(30), 5128–5147 (2012).
- Morley AD, Pugliese A, Birchall K *et al.* Fragment-based hit identification: thinking in 3D. *Drug Discov. Today.* 18(23–24), 1221–1227 (2013).
- Mok NY, Brenk R, Brown N. Increasing the coverage of medicinal chemistry-relevant space in commercial fragments screening. *J. Chem. Inf. Model.* 54(1), 79–85 (2014).
- Aldeghi M, Malhotra S, Selwood DL, Chan AWE. Two- and three-dimensional rings in drugs. *Chem. Biol. Drug Des.* 83(4), 450–461 (2014).
- **Analysis of shape profiles of rings in drugs.**
- Kombo DC, Tallapragada K, Jain R *et al.* 3D molecular descriptors important for clinical success. *J. Chem. Inf. Model.* 53(2), 327–342 (2013).
- **Evidence of key benefits to increased three-dimensionality in drug-like molecules.**
- Lovering F, Bikker J, Humblet C. Escape from flatland: increasing saturation as an approach to improving clinical success. *J. Med. Chem.* 52(21), 6752–6756 (2009).
- Lovering F. Escape from flatland 2: complexity and promiscuity. *MedChemComm* 4(3), 515–519 (2013).
- Kortagere S, Krasowski MD, Ekins S. The importance of discerning shape in molecular pharmacology. *Trends Pharmacol. Sci.* 30(3), 138–147 (2009).
- Over B, Wetzel S, Grütter C *et al.* Natural-product-derived fragments for fragment-based ligand discovery. *Nat. Chem.* 5(1), 21–28 (2013).
- Hewings DS, Wang M, Philpott M *et al.* 3,5-Dimethylisoxazoles act as acetyl-lysine-mimetic bromodomain ligands. *J. Med. Chem.* 54(19), 6761–6770 (2011).
- Christ F, Voet A, Marchand A *et al.* Rational design of small-molecule inhibitors of the LEDGF/p75-integrase interaction and HIV replication. *Nat. Chem. Biol.* 6(6), 442–448 (2010).
- Bodur C, Basaga H. Bcl-2 inhibitors: emerging drugs in cancer therapy. *Curr. Med. Chem.* 19(12), 1804–1820 (2012).
- Zhao Y, Aguilar A, Bernard D, Wang S. Small-molecule inhibitors of the MDM2-p53 protein–protein interaction (MDM2 inhibitors) in clinical trials for cancer treatment. *J. Med. Chem.* 58(3), 1038–1052 (2015).
- Yang Y, Engkvist O, Llinàs A, Chen H. Beyond size, ionization state, and lipophilicity: influence of molecular topology on absorption, distribution, metabolism, excretion, and toxicity for druglike compounds. *J. Med. Chem.* 55(8), 3667–3677 (2012).
- Ishikawa M, Hashimoto Y. Improvement in aqueous solubility in small molecule drug discovery programs by disruption of molecular planarity and symmetry. *J. Med. Chem.* 54(6), 1539–1554 (2011).
- Todeschini R, Consonni V. *Molecular Descriptors for Chemoinformatics, Volume 41 (2 Volume Set)*. John Wiley & Sons, Weinheim, Germany (2009).
- Sauer W, Schwarz MK. Molecular shape diversity of combinatorial libraries: a prerequisite for broad bioactivity. *J. Chem. Inf. Comput. Sci.* 43(3), 987–1003 (2003).
- **Comprehensive description of the principal moments of inertia.**
- Firth NC, Brown N, Blagg J. Plane of best fit: a novel method to characterize the three-dimensionality of molecules. *J. Chem. Inf. Model.* 52(10), 2516–2525 (2012).
- **Comprehensive description of the plane of best fit.**
- Gaulton A, Bellis LJ, Bento AP *et al.* ChEMBL: a large-scale bioactivity database for drug discovery. *Nucl. Acids Res.* 40(D1), D1100–D1107 (2011).
- Schuffenhauer A, Ertl P, Roggo S, Wetzel S, Koch MA, Waldmann H. The scaffold tree – visualization of the scaffold universe by hierarchical scaffold classification. *J. Chem. Inf. Model.* 47(1), 47–58 (2007).
- Firth NC, Atrash B, Brown N, Blagg J. MOARF, an integrated workflow for multiobjective optimization: implementation, synthesis, and biological evaluation. *J. Chem. Inf. Model.* 55(6), 1169–1180 (2015).
- Schwab CH. Conformations and 3D pharmacophore searching. *Drug Discov. Today Technol.* 7(4), e245–e253 (2010).
- Perola E, Charifson PS. Conformational analysis of drug-like molecules bound to proteins: an extensive study of ligand reorganization upon binding. *J. Med. Chem.* 47(10), 2499–2510 (2004).
- Sadowski J, Gasteiger J, Klebe G. Comparison of automatic three-dimensional model builders using 639 x-ray structures. *J. Chem. Inf. Model.* 34(4), 1000–1008 (1994).

- 26 ChEMBL v21.
www.ebi.ac.uk/chembl/downloads
- 27 Lipinski CA, Lombardo F, Dominy BW, Feeney PJ. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23(1–3), 3–25 (1997).
- 28 Mallinson J, Collins I. Macrocycles in new drug discovery. *Future Med. Chem.* 4(11), 1409–1438 (2012).
- 29 Accelrys. Pipeline Pilot version 9.5.0.831 available from BIOVIA.
<http://accelrys.com>
- 30 CORINA, version 3.4, Molecular Networks GmbH: Erlangen, Germany (2013).
- 31 Gasteiger J, Rudolph C, Sadowski J. Automatic generation of 3D-atomic coordinates for organic molecules. *Tetrahedron Computer Methodology* 3(6), 537–547 (1990).
- 32 RDKit: Open-source cheminformatics.
www.rdkit.org
- 33 Group CC. Molecular Operating Environment 2015.10 available from Chemical Computing Group.
<http://chemcomp.com>
- 34 Wirth M, Volkamer A, Zoete V *et al.* Protein pocket and ligand shape comparison and its application in virtual screening. *J. Comput. Aided Mol. Des.* 27(6), 511–524 (2013).
- 35 Harper M, Weinstein B, Simon C *et al.* Python-ternary: ternary plots in Python. Zenodo.
<http://zenodo.org/record/34938#.V8AmUsWDpzU>
- 36 Brown DG, Boström J. Analysis of past and present synthetic methodologies on medicinal chemistry: where have all the new reactions gone? *J. Med. Chem.* 59(10), 4443–4458 (2016).
- 37 Langdon SR, Brown N, Blagg J. Scaffold diversity of exemplified medicinal chemistry space. *J. Chem. Inf. Model.* 51(9), 2174–2185 (2011).
- 38 Hann MM, Leach AR, Harper G. Molecular complexity and its impact on the probability of finding leads for drug discovery. *J. Chem. Inf. Comput. Sci.* 41(3), 856–864 (2001).